J Sign Process Syst (2014) 74:19–31
DOI 10.1007/s11265-013-0825-4

Multimedia Event Detection Using Segment-Based Approach for Motion Feature

Sang Phan · Thanh Duc Ngo · Vu Lam · Son Tran · Duy-Dinh Le · Duc Anh Duong · Shin'ichi Satoh

Received: 21 January 2013 / Revised: 29 June 2013 / Accepted: 1 July 2013 / Published online: 30 July 2013

Abstract Multimedia event detection has become a popular research topic due to the explosive growth of video data. The motion features in a video are often used to detect events because an event may contain some specific actions or moving patterns. Raw motion features are extracted from the entire video first and then aggregated to form the final video representation. However, this video-based representation approach is ineffective when used for realistic videos because the video length can be very different and the clues for determining an event may happen in only a small segment of the entire video. In this paper, we propose using a segment-based approach for video representation. Basically, original videos are divided into segments for feature extraction and classification, while still keeping the evaluation at the video level. The experimental results on recent TRECVID Multimedia Event Detection datasets proved the effectiveness of our approach.

Keywords Multimedia event detection · Segment-based · Video-based · Dense trajectories

S. Phan (✉) · T. D. Ngo
The Graduate University for Advanced Studies (SOKENDAI), Shonan Village, Hayama, Kanagawa 240-0193, Japan
e-mail: plsang@nii.ac.jp

T. D. Ngo
e-mail: ndthanh@nii.ac.jp

V. Lam · S. Tran
Faculty of Information Technology, Ho Chi Minh City University of Science, 227 Nguyen Van Cu St., Dist. 5, Ho Chi Minh City, Vietnam

V. Lam
e-mail: lqvu@fit.hcmus.edu.vn

S. Tran
e-mail: ttson@fit.hcmus.edu.vn

D.-D. Le · D. A. Duong
Multimedia Communications Lab, University of Information Technology, KM20 Xa Lo Ha Noi, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

D.-D. Le
e-mail: duyld@uit.edu.vn

D. A. Duong
e-mail: ducda@uit.edu.vn

D.-D. Le · S. Satoh
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

S. Satoh
e-mail: satoh@nii.ac.jp

1 Introduction

Multimedia Event Detection (MED) is a challenging task in TREC Video Retrieval Evaluation (TRECVID).[1] The task is defined as follow: given a collection of test videos and a list of test events, indicate whether each of the test events is present in each of the test videos. The aim of MED is to develop systems that can automatically find video containing any event of interest, assuming only a limited number of training exemplars are given.

The need for such MED systems is rising because a massive number of videos are produced every day. For example, more than 3 million hours of video are uploaded and over 3 billion hours of video are watched each month on YouTube,[2] the most popular video sharing website. What is

[1] http://trecvid.nist.gov/

[2] http://www.youtube.com/t/press_statistics

needed are the tools for automatically processing the video content and looking for the presence of a complex event in such unconstrained capturing videos. Automatic detection of complex events has great potential for many applications in the field of web video indexing and retrieval. In practice, a viewer may only want to watch goal scenes in a long football video, a housewife may need to search for videos that teach her how to make a cake, a handyman may look for how to repair an appliance, or a TV program manager may want to remove violent scenes in a film before it is aired.

However, detecting events in multimedia videos is a difficult task due to both the large content variation and uncontrolled capturing conditions. The video content is extremely diverse even in a same event class. The genres of video are also very varied, such as interviews, home videos, and tutorials. Moreover, the number of events is expected to be extensive for large scale processing. Each event, in its turn, can involve a number of objects and actions in a particular setting (indoors, outdoors, etc). Furthermore, multimedia videos are typically recorded under uncontrolled conditions such as different lighting, viewpoints, occlusions, complicated camera motions and cinematic effects. Therefore, it is very hard to model and detect of multimedia events.

The most straightforward approach toward building a large scale event detection system is using a bag-of-words (BoW) model [2]. There are two types of BoW representations that are used for MED: BoW representation at the keyframe level and BoW representation at the video level. The first method is employed for still image features where the keyframes are often extracted at a fixed interval. The second method is employed for motion features where moving patterns from the entire video are extracted. These methods are respectively referred to as keyframe-based [8, 10, 17] and video-based [8, 10] in this paper. Although these methods can obtain reasonable results, they all suffer from severe limitations. For the keyframe-based approach, temporal information is not incorporated in the model. Moreover, it is possible that important keyframes are missed extraction. Extracting more keyframes can tackle this problem but the scalability is also a problem for concern. On the other hand, the video-based approach is most likely to suffer from noise. We found that the video length is very different from video to video (even from videos of the same event class). In addition, the clues to determine an event may appear within a small segment of the entire video. Thus, comparing the BoW representation of two videos is unreliable because it may contain unrelated information. Figure 1 illustrates these limitations for both approaches.

In this paper, we propose using a segment-based approach to overcome the limitations of both the keyframe-based and video-based approaches. The basic idea is to examine shorter segments instead of using the representative frames or entire video. We can reduce the amount of unrelated information in the final representation, while still benefiting from the temporal information by dividing a video into segments. In particular, we investigate two methods to cut a video into segments. The first method is called uniform sampling, where every segment has an equal length. We choose different segment lengths and use two types of sampling: non-overlapping and overlapping. The overlapped configuration is used to test the influence of dense segment sampling. The second method divides the video based on the shot boundary detection to take into account the boundary information of each segment. Once segments are extracted, we use dense trajectories, a state-of-the-art motion feature proposed by Wang [23], for the feature extraction. After that, a BoW model is employed for the feature representation. The experimental results on TRECVID MED 2010 and TRECVID MED 2011 showed the improvement of the segment-based approach over the video-based approach. Moreover, a better performance can be obtained by using the overlapping sampling strategy.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 gives an overview of the dense trajectory motion feature and our segment-based approach. The experimental setup including an introduction to the benchmark dataset and the evaluation method are presented in Section 4. Then, in Section 5, we present and analyze our experimental results. Detailed discussions of these results are presented in Section 6. Finally, Section 7 concludes the paper with discussions on our future work.

2 Related Work

Challenges began from TRECVID 2010,[3] and Multimedia Event Detection has drawn the attention of many researchers. Seven teams participated in the debut challenge and 19 teams participated the following year (MED 2011). Many MED systems have been built and different strategies have been used for the event detection system.

Columbia University (CU) team achieved the best result in TRECVID MED 2010. Their success greatly influenced later MED systems. In their paper [10], they answered two important questions. The first question was, “What kind of feature is more effective for multimedia event detection?”. The second one was, “Are features from different feature modalities (e.g., audio and visual) complementary for event detection?”. Different kinds of features have been studied, such as SIFT [16] for the image feature, STIP [14] for the motion feature and MFCC (Mel-frequency cepstral coefficients [15]) for the audio feature to answer the first question. In general, the STIP motion feature is the best single feature

[3] www.nist.gov/itl/iad/mig/med10.cfm

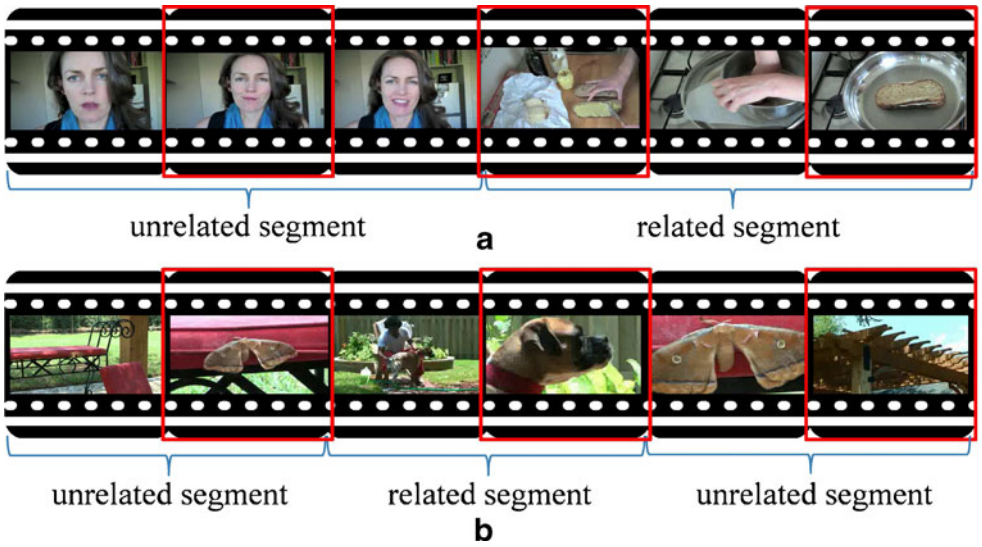


Figure 1 **a** Example video for "making a sandwich" event: the related segment appears after a self-cam segment (unrelated); **b** example video for "grooming an animal" event: related segment is sandwiched between two unrelated segments. This kind of video is popular in realistic video datasets like MED. The frames with a red outlined box are examples of the extracted keyframes when using a keyframe-based approach, which suffers from both noise and missed extraction.

for MED. However, the system should combine strong complementary features from multiple modalities (both visual and audio) in order to achieve better results.

The IBM team [8] achieved the runner-up MED system in TRECVID 2010. They incorporated information from a wide range of static and dynamic visual features to build their baseline detection system. For the static features, they used the local SIFT [16], GIST [20] descriptors and various global features such as Color Histogram, Color Correlogram, Color Moments, Wavelet Texture, etc. They used the STIP [14] feature with a combined HOG-HOF [13] descriptor for the dynamic feature.

The Nikon MED 2010 system [17] is also a remarkable system due to its simple but effective solution. They built a MED system based on the assumption that a small number of images in a given video contain enough information for event detection. Thus, they reduced the event detection task to the classification problem for a set of images, called keyframes. However, keyframe extraction is based on a scene cut detection technique [7] that is less reliable in realistic videos. Moreover, the scene length is not consistent, which may affect the detection performance.

The BBN Viser system [18] achieved the best performance at TRECVID MED 2011. Their success confirmed the effectiveness of the multiple modalities approach for multimedia event detection. In their work, they further investigated the performance of the appearance features (e.g., SIFT [16]), color feature (e.g. RGB-SIFT [22]), and motion (e.g., STIP [14]), and also MFCC [15] based audio features. Different kinds of fusion strategies have been explored, from which the novel non-parametric fusion strategy based on a video specific weighted average fusion has shown promising results.

In general, most systems used the multiple modalities approach to exploit different visual cues to build their baseline detection systems. Static image characteristics are extracted from frames within provided videos. Colombia University's results [10] suggest that methods for exploiting semantic content from web images, such as [5] and [10], are not effective for multimedia event detection. For motion characteristics, most systems employed the popular STIP proposed by Laptev in [14] for detecting complex actions. Other systems also took into account the HOG3D [12] and MoSIFT [1] motion features. All these systems used a video-based approach for the motion features, i.e., the motion features are extracted from the entire video. IBM's MED system [8] also applied the video-based approach but the video was downsampled to five frames per second. One drawback of this video-based approach is that it may encode unrelated information in the final video representation. In a long video, the event information may happen during a small segment, and the information from the other segments tends to be noisy. That is why it is important to localize the event segment (i.e., where the event happens). This problem has been thoroughly investigated by Yuan et al. [25]. Yuan proposed using a spatio-temporal branch-and-bound search to quickly localize the volume where an action might happen. In [24], Xu proposed a method to find optimal frame alignment in the temporal dimension to recognize events in broadcast news. In [6], a transfer learning method is proposed to recognize simple action events. However, these works are not applicable for complex actions in multimedia event videos.

Different from other approaches, we use a segment-based approach for the event detection. We did not try to localize the event volume like Yuan in [25]. In a simpler way, we use a uniform sampling with different segment lengths for our evaluation. We also investigate the benefit of using the shot boundary detection technique in [7] for dividing video into segments. Moreover, an overlapped segment sampling

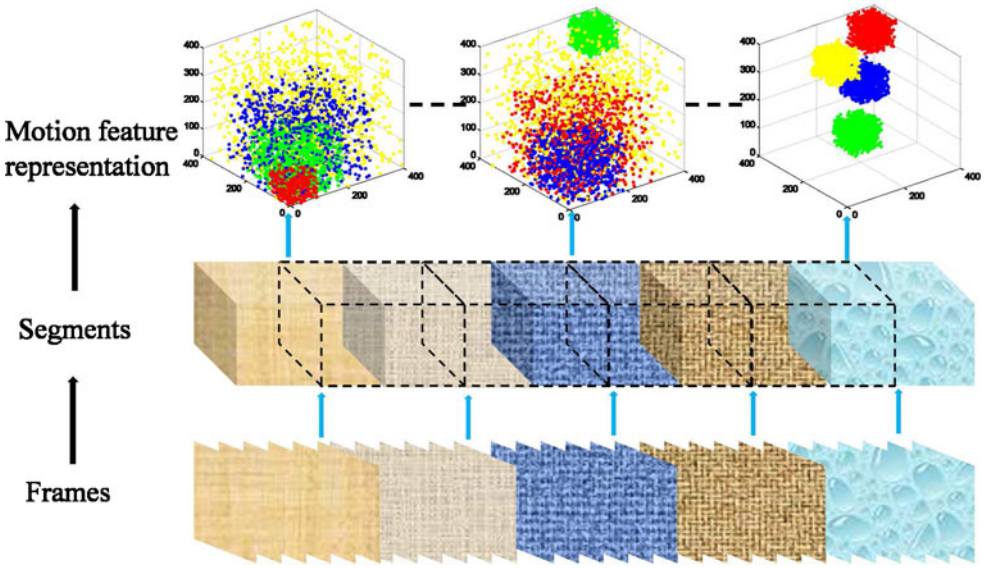


Figure 2 Illustration of our segment-based approach. The original video is divided into segments by using non-overlapping and overlapping sampling (overlapped segment examples are drawn in *dashes*). After that, the feature representation is separately calculated for each segment. This figure is best viewed in color.

strategy is also considered for a denser sampling. To the best of our knowledge, no MED system has previously used this approach. We evaluate its performance using the dense trajectories motion feature that was recently proposed by Wang in [23]. The dense trajectories feature has achieved state-of-the-art performances for various video datasets, including challenging datasets like Youtube Action[4] and UCF Sports.[5] In TRECVID MED 2012, the dense trajectories feature was also widely used by top performance systems such as AXES [21], and BBNVISER [19]. We use the popular "bag-of-words" model in [2] as our feature representation technique. Finally, we use a Support Vector Machine (SVM) classifier for the training and testing steps.

3 Dense Trajectories and Segment-Based Approach

We introduce the dense trajectory motion feature proposed by Wang in [23] in this section. We additionally briefly review the trajectory extraction and description method. A detailed calculation of all the related feature descriptors, especially for Motion Boundary Histogram, is also presented. Our segment-based approach for motion features is introduced at the end of this section.

3.1 Dense Trajectories

Trajectories are obtained by tracking the densely sampled points using the optical flow fields. First, the feature points are sampled on a grid with a step size of 5 pixels and at multiple scales spaced by a factor of $1/\sqrt{2}$. Then, the feature points are separately tracked in each scale. Each point $P_t = (x_t, y_t)$ at frame t is tracked to the next frame t+1 by using median filtering in a dense optical flow field $\omega = (u_t, v_t)$:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)}, \quad (1)$$

where M is the median filter, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t).

After extracting a trajectory, two kinds of feature descriptors are adopted: a trajectory shape descriptor and a trajectory-aligned descriptor.

Trajectory Shape Descriptor The trajectory shape descriptor is the simplest one for representing an extracted trajectory. It is defined based on the displacement vectors. Given a trajectory of length L, its shape is described by the sequence $S = (\Delta P_t, ..., \Delta P_{t+L-1})$, where $\Delta P_t = P_{t+1} - P_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. The resulting vector is then normalized by the sum of the magnitudes of the displacement vectors:

$$S' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \quad (2)$$

Trajectory-aligned Descriptor More complex descriptors can be computed within a space-time volume around the trajectory. The size of the volume is N × N spatial pixels and L temporal frames. This volume is further divided into a $n_\sigma \times n_\sigma \times n_\tau$ grid to encode the spatial-temporal information between the features. The default settings for these parameters are N = 32 pixels, L = 15 frames, $n_\sigma = 2$, and $n_\tau = 3$. The features are separately calculated and aggregated in

[4] http://www.cs.ucf.edu/~liujg/YouTube_Action_dataset.html
[5] http://www.cs.ucf.edu/vision/public_html

Table 1 List of events and its number of positive samples in event collection set of MED 2011 dataset.

Event Id	Event name	#Pos videos
E001	Attempting a board trick	173
E002	Feeding an animal	168
E003	Landing a fish	152
E004	Wedding ceremony	163
E005	Working on a woodworking project	159
E006	Birthday party	221
E007	Changing a vehicle tire	119
E008	Flashmob gathering	191
E009	Getting a vehicle unstuck	151
E010	Grooming an animal	143
E011	Making a sandwich	186
E012	Parade	171
E013	Parkour	134
E014	Repairing an appliance	137
E015	Working on a sewing project	124

each region. Finally, the features in all regions are concatenated to form a single representation for the trajectory. Three kinds of descriptors have been employed for representing trajectory following this design: The Histogram of Oriented Gradient (HOG), which was proposed by Dalal et al. in [4] for object detection, The Histogram of Optical Flow (HOF), which was used by Laptev in [13] for human action recognition, and the Motion Boundary Histogram (MBH). The MBH descriptor was also proposed by Dalal et al. [3] for human detection, where the derivatives are computed separately for the horizontal and vertical components of the optical flow $I_\omega = (I_x, I_y)$. The spatial derivatives are computed for each component of the optical flow field I_x and I_y independently. After that, the orientation information is quantized into histogram, similarly to that for the HOG descriptor (8-bin histogram for each component). Finally, these two histograms are normalized separately with the L_2 norm and concatenated together to form the final representation. Since the MBH represents the gradient of the optical flow, constant motion information is suppressed and only the information concerning the changes in the flow field (i.e., motion boundaries) is kept.

According to the author [23], the MBH descriptor is the best feature descriptor for dense trajectories. One interesting property of the MBH is that it can cancel out camera motion. That is why it shows significant improvement on realistic action recognition dataset compared to other trajectory descriptors. We only use the MBH descriptor in this study to test the performance of our proposed segment-based method.

3.2 Segment-Based Approach for Motion Feature

Our proposed segment-based approach is as follows. At first, the video is divided into fixed length segments. We choose different segment lengths to pick the optimal one. In particular, we choose segment lengths of 30, 60, 90, 120, 200 and 400 s. The lengths of 120 and 60 s are respectively close to the mean (115 s) and geometric mean (72 s)

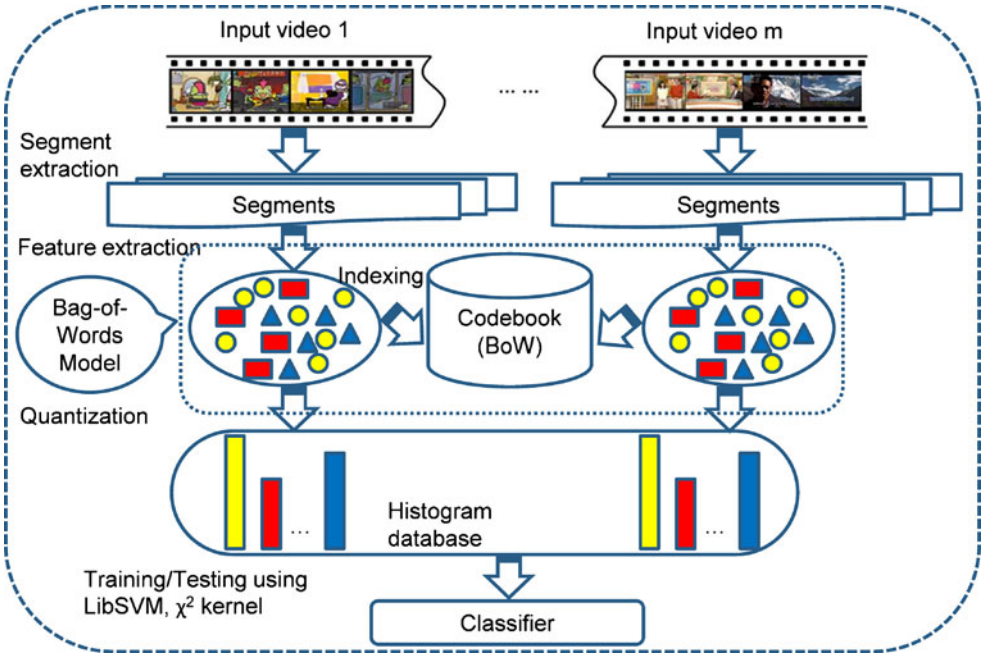


Figure 3 Evaluation framework for our baseline MED system.

Table 2 Results on the MED 2010 dataset using non-overlapping sampling.

Event/MAP	30 s	60 s	90 s	120 s	200 s	400 s	Late fusion
Assembling a shelter	0.4140	0.4511	0.4339	0.4457	0.4595	**0.4610**	0.4532
Batting in a run	0.7650	0.7852	0.7799	0.7553	0.7823	**0.7871**	0.7181
Making a cake	0.3596	0.3636	0.3433	0.3569	0.3058	0.3032	**0.3727**
All	0.5129	**0.5333**	0.5190	0.5193	0.5158	0.5171	0.5146

length of the training dataset. The geometric mean value is also considered because it can eliminate the influence of outline cases, i.e., videos of exceptionally long durations. After that, the dense trajectory features are extracted from the entire segment. A "bag-of-words" approach is used to generate the final representation for each segment from the raw trajectory features (Fig. 2).

For the previous segment-based approach, a video is divided into continuous segments. This means information about the semantic boundary of a segment is not taken into account. However, this information is important because it keeps the semantic meaning of each segment. The simplest way to overcome this drawback is to use a denser sampling such as the overlapped segments. We use an overlapping strategy for the same segment length as in the non-overlapping experiments. In practice, we use uniform segment sampling with 50 % of overlapping. This means the number of segments will be doubled for each overlapping experiment.

Another way to extract segments with boundary information is to employ a shot boundary detection technique. For a fast implementation, we use the algorithm proposed in [7]. This technique is also used in the Nikon 2010 MED system [17]. Basically, at first, this method constructs a space-time image from the input video. We can sample points or calculate the color histogram for each frame to construct the space-time image. This will reduce the 2D frame image to the space dimension of the space-time image. The time dimension is the number of frames of the video. The Canny edge detection algorithm is used to detect the vertical lines after attaining the space-time image. Each detected vertical line is considered as a scene cut. The method in [7] also proposed solutions for other kinds of scene transitions such as a fade or wide. However, from our previous study, this method showed poor results in these cases. Thus, we only adopted the scene cut detection algorithm. Each detected scene cut is considered a segment in our experiments.

Our proposed segment-based approach is compared with the video-based one. Actually, when the segment length is long enough, it becomes the entire video. In that case, we can consider the video-based approach a special type of segment-based approach.

4 Experimental Setup

4.1 Dataset

We tested our method on TRECVID MED 2010 and TRECVID MED 2011 datasets. An event kit is provided with the definitions and textual descriptions for all the events for each dataset. The first dataset contains 3,468 videos, including 1,744 videos for training and 1,724 video clips for testing, containing a total of more than 110 video hours. In TRECVID MED 2010, there are 3 events classes: assembling a shelter, batting in a run, and making a cake. The TRECVID MED 2011 dataset defined the 15 event classes listed in Table 1. The first five events (E001-E005) are used for training and validation and the last 10 events (E006-E015) are used for testing. It comprises of over 45,000 video clips for a total of 1,400 h of video data. All the video clips are divided into three sets: event collection (2392 video clips), development collection (10198 video clips), and test collection (31,800 video clips). It is worth noting that these two datasets contain a major number of background video clips, i.e., video clips that do not belong to any event. The number of positive videos in the event collection is also listed in Table 1.

4.2 Evaluation Method

Figure 3 shows our evaluation framework for the motion features. We conducted experiments using the proposed segment-based approach and the video-based approach for comparison. We use the library published online by the author[6] to extract dense trajectory feature. The source code is customized for pipeline processing using only an MBH descriptor to save computing time but other parameters are set to default. Due to the large number of features produced when using the dense sampling strategy, we use the "bag-of-words" approach to generate the features for each segment. At first, we randomly select 1,000,000 dense trajectories for clustering to form a codebook of 4000 visual codewords.

[6] http://lear.inrialpes.fr/people/wang/dense_trajectories

Table 3 Results on the MED 2010 dataset using overlapping sampling.

Event/MAP	30 s	60 s	90 s	120 s	200 s	400 s	Late fusion
Assembling a shelter	0.4177	**0.4781**	0.4617	0.4614	0.4601	0.4682	0.4486
Batting in a run	0.7727	0.7918	**0.7975**	0.7886	0.7893	0.7756	0.7691
Making a cake	0.4083	0.3819	0.3155	0.3415	0.3464	0.3239	**0.4232**
All	0.5329	**0.5506**	0.5249	0.5305	0.5319	0.5226	0.5470

After that, the frequency histogram of the visual words is computed over the videos/segments to generate the final feature vector. We also adopt the soft assignment weighting scheme, which was initially proposed by Jiang in [9], to improve the performance of the "bag-of-words" approach.

Once all the features are extracted, we use the popular Support Vector Machine (SVM) for the classification. In particular, we use the LibSVM library available online[7] and adopt the one-vs.-rest scheme for multi-class classification. We annotate the data in the following way to prepare it for the classifier. All the videos/segments from positive videos are considered positive samples, and the remaining videos/segments (in the development set) are chosen as the negative samples. For testing purposes, we also use the LibSVM to predict the scores of the videos/segments in each testing video. The score of a video is defined as the largest score among its videos/segments. This score indicates how likely a video belongs to an event class.

5 Experimental Results

This section presents the experimental results from using our proposed approach on the MED 2010 and MED 2011 dataset. We also present the results of combining various segment lengths using the late fusion technique. This is a simple fusion technique where the predicted score of each video is the average one of that video in all combined runs. We also report the performance of our baseline event detection system using the keyframe-based and video-based approach for comparison.

All the experiments were performed on our grid computers. We utilized up to 252 cores for the parallel processing using Matlab codes. All the results are reported in terms of the Mean Average Precision (MAP). We calculate MAP using the TRECVID evaluation tool[8] from the final score of each video in the test set. The best performing feature is highlighted in bold for each event.

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[8] http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/

5.1 On TRECVID MED 2010

5.1.1 Non-Overlapping and Overlapping Sampling

Table 2 lists the results from our segment-based approach when using a non-overlapping sampling strategy. These results show that the performance is rather sensitive to the segment length and it is also event-dependent. For example, the detection results of the first event, "assembling a shelter", are better when the segment length is increased. On the other hand, the "making a cake" event tends to be more localized, i.e. the shorter the segment, the better the performance. The performance of the "batting in a run" event is quite stable when segment length is longer than 60 s. However, it is decreased 2 % at 30 s. This suggests that shorter lengths can harm the performance. In general, the performance of a 60-s segment is the best. This length is also around the geometric mean length of the training set. Thus, we got peak results for segment length around geometric mean point.

We further investigated the performance of a denser segment sampling, i.e., an overlapping sampling strategy. Interestingly, the MAP score in Table 3 is consistently increased for each event compared to the results without using overlapped segments. Figure 4 shows a detailed comparison between the two strategies in terms of the over-all performance. We again found that the performance with a segment length around the geometric mean length (60 s) was the best. We also combined the performances of all the segment lengths using late fusion and the results are listed in the last column of Tables 2 and 3. The late fusion strategy can benefit the "making a cake" event, but it decreased the performances of the remaining events. The overall performance is lower than the best one.

5.1.2 Segment Sampling Based on Shot Boundary Detection

The second column in Table 4 shows the performance when shot boundary detection is used to extract segments. Unexpectedly, the performance is quite low even when compared with the video-based approach (listed in the last column). There are two possible reasons for this low level of performance: (1) The shot boundary detection technique is

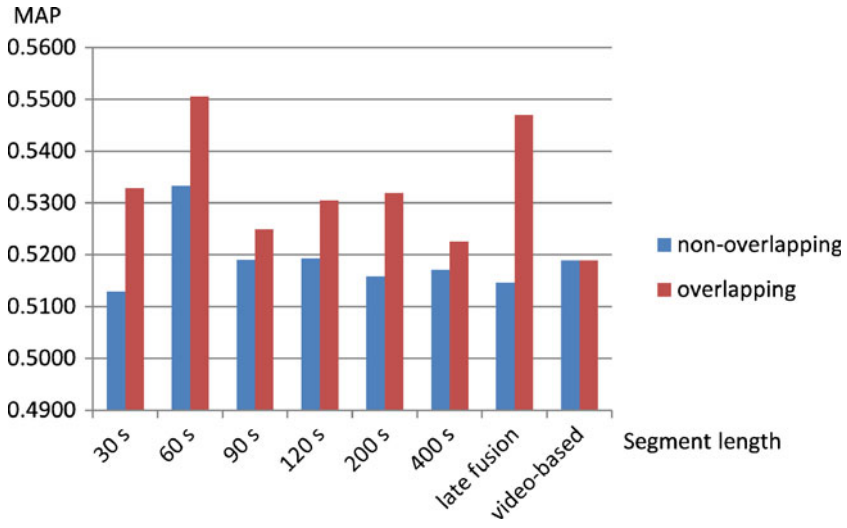


Figure 4 Results from using segment-based approach with non-overlapping and overlapping sampling on MED 2010. In all cases, the overlapping sampling performs the best.

inaccurate when used on uncontrolled capturing videos; (2) the shot units may not contain enough information to determine an event. The second reason suggests that combining multiple shots to form a segment may improve the performance. Thus, we have conducted a segment-based experiment based on this observation using segments extracted from multiple shots. However, we did not see any significant improvement. Thus, the first reason is why this experiment had poor result.

We also included the best results from the segment-based experiments using non-overlapping and overlapping sampling in Table 4 for comparison. In general, our segment-based approach outperforms the video-based approach by more than 3 % in terms of MAP. We did not conduct a keyframe-based experiment because we learned that it is inefficient compared to the video-based approach.

5.2 On TRECVID MED 2011

We conducted the same segment-based experiments on MED 2011. For both the non-overlapping and overlapping experiments, we chose segment lengths of 60, 90, 120, and 200 s and compare them with the video-based approach. A late fusion strategy is also used to combine the performances of different segment lengths. We did not conduct a shot boundary detection experiment because we showed that it is inefficient. Tables 5 and 6 list the performances of each event for non-overlapping and overlapping experiment, respectively. Figure 5 shows a better view for comparing the overall performance. The result from using video-based approach, which is 0.2095 MAP, is also included for comparison. In most cases, the overlapping sampling had better results than the non-overlapping sampling. In all cases, the segment-based approach also outperforms the video-based approach. The best improvement was about 5 %, which was obtained at 120 s using an overlapping sampling. The late fusion run also confirms its effectiveness for some events, such as "Flash-mob gathering" and "Working on a sewing project".

6 Discussions

6.1 Optimal Segment Length

It is true that the lengths of the event segments are quite different, even for the same events. Therefore, the fixed length video segments are obviously not the optimal solution to describe the events. However, compared to the video-based approach, as shown in our experiments on the datasets of TRECVID MED 2010 and TRECVID MED 2011, the segment-based approach using overlapping strategy for extracting segments consistently outperforms.

Table 4 Comparison of different segment-based approaches with the video-based approach on the MED 2010 dataset.

Event/MAP	Best non-overlapping	Best overlapping	SBD segments	Video-based
Assembling shelter	0.4511	0.4781	0.4284	**0.4911**
Batting in a run	0.7852	**0.7918**	0.7866	0.7902
Making a cake	0.3636	**0.3819**	0.1918	0.2755
All	0.5333	**0.5506**	0.4689	0.5189

Table 5 Results on the MED 2011 dataset using non-overlapping sampling.

Event/MAP	60 s	90 s	120 s	200 s	Late fusion
E006	0.1060	**0.1277**	0.1162	0.1005	0.1217
E007	0.1003	**0.1521**	0.1461	0.0539	0.1419
E008	0.4811	0.4923	0.4840	0.4508	**0.4975**
E009	0.2077	0.2072	0.1962	0.1860	**0.2145**
E010	0.0794	**0.0916**	0.0486	0.0854	0.0771
E011	**0.0943**	0.0698	0.0903	0.0703	0.0805
E012	0.3061	0.3560	0.3052	**0.3639**	0.3309
E013	0.5974	0.6030	0.5861	0.5941	**0.6033**
E014	0.2307	0.2008	**0.2772**	0.1723	0.2585
E015	0.1364	**0.1599**	0.1357	0.1284	0.1583
All	0.2340	0.2460	0.2386	0.2206	**0.2484**

It is ideal if the boundary of the event segment can be determined. However, this localization problem is difficult. The straightforward way to tackle this problem is extracting segments based on shot boundary information. This solution is reasonable because the event might be localized in certain shots. However, we obtained unexpected results due to the unreliability of shot boundary detection in uncontrolled video dataset and the event segment might span to several shots.

The method described in [11] suggests another approach to divide a video into segments. Instead of learning a randomized spatial partition for images, we can learn a randomized temporal partition for videos. However, this approach needs sufficient positive training samples while MED datasets have a small number of positive samples with large variation. On the other hand, it is also not scalable because learning and testing the best randomized pattern is time-consuming. Therefore, the fixed-length approach is quite simple but still effective.

Supposed the segment length is fixed, what is the optimal segment length for event detection? This is a difficult question and the answer depends on the dataset. The results of late fusion are quite close to the peak performance of each experiment. This suggests a methodical way to choose the optimal segment length, i.e., combining multiple lengths together (which is similar to [11]). However, to achieve the scalability, we should reduce the number of combined lengths as much as possible. From the experimental results on both the MED 2010 and MED 2011 dataset, we observed that with segment length from 60 s to 120 s, the performance is rather stable and close to the peak result. Interestingly, this range is approximate to the range from the geometric mean length to (arithmetic) mean length of the training sets. We also combined multiple segment lengths together using late fusion with equal weights for all segment lengths for comparison. There are two combined runs: one for segment lengths from 60 s to 120 s and the other is for all segment lengths. The result obtained when combining segment lengths from 60 s to 120 s is equivalent to the result obtained when combining all lengths, as shown in Table 7. Therefore, based on this observation, we can choose the first combined run as an efficient way for solving the optimal segment length problem of the proposed segment-based approach on other datasets.

Table 6 Results on the MED 2011 dataset using overlapping sampling.

Event/MAP	60 s	90 s	120 s	200 s	Late fusion
E006	0.1074	0.1069	**0.1151**	0.1010	0.1086
E007	0.1570	**0.1733**	0.1552	0.1466	0.1610
E008	0.4788	0.4767	0.4969	0.4620	**0.4903**
E009	0.1830	0.1999	**0.2160**	0.1972	0.1954
E010	**0.1150**	0.0851	0.1008	0.0746	0.1108
E011	0.0602	0.0885	**0.1591**	0.0779	0.0819
E012	**0.3674**	0.3129	0.3150	0.3075	0.3293
E013	0.6025	0.5893	**0.6188**	0.5675	0.5872
E014	0.2718	0.2487	**0.2744**	0.2095	0.2706
E015	0.1777	0.1459	0.1562	0.1214	**0.1795**
All	0.2521	0.2427	**0.2607**	0.2265	0.2515

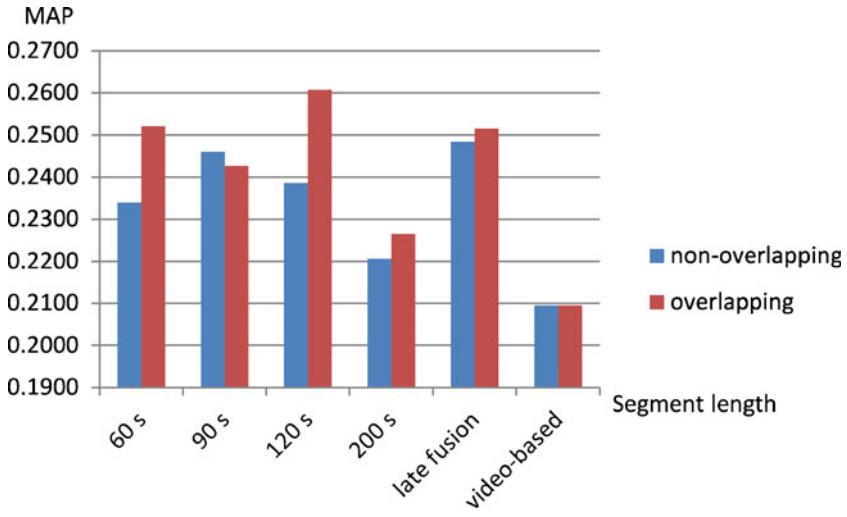


Figure 5 Results from using segment-based approach with non-overlapping and overlapping sampling on MED 2011. In most cases, the overlapping sampling performs the best.

6.2 Scalability

For scalability, we discuss the storage and computation costs of our experiments. At first, our system does not consume a lot of disk storage because we only store the final representation of the videos or segments, not the raw features. We calculated the BoW features directly from the raw feature outputs using a pipeline reading technique. One drawback is that this technique requires a lot of memories. However, we handled this problem by encoding the raw features into smaller chunks and aggregating them to generate the final representation. By this way, we can manage the mount of memory usage.

In our framework, the most time-consuming steps are the feature extraction and representation (using the bag-of-words model). It is worth noting that the computation time for one video is independent of the segment length, which means our segment-based approach has the same computational cost as the video-based approach. On the other hand, when we do experiments at the segment level, we will have more training and testing samples than that in the video-based approach. Thus, it will cost more in time to train and test using the segment-based approach. However, this cost is relatively small compared with the feature extraction and representation cost. For example, when using a grid computer with 252 cores, it took us about 10 h to generate the feature representation for each segment-based experiment on MED 2010 dataset. In the mean time, we used one-core processor for the training and testing, but it only took about 4–8 h for the training and 2–4 h for the testing on each event. For the MED 2011 dataset, the computational cost was around 13 times bigger than the MED 2010 (linearly to the number of videos it contains).

Table 7 Comparison of different segment-based approaches with the video-based approach on the MED 2011 dataset.

Event/MAP	Non-overlapping sampling			Overlapping sampling			Video-based
	Best (at 90 s)	Late fusion (all lengths)	Late fusion (60, 90, 120 s)	Best (at 120 s)	Late fusion (all lengths)	Late fusion (60, 90, 120 s)	
E006	**0.1277**	0.1217	0.1244	0.1151	0.1086	0.1083	0.0959
E007	0.1521	0.1419	0.1369	0.1552	**0.1610**	0.1616	0.1303
E008	0.4923	**0.4975**	0.4973	0.4969	0.4903	0.4871	0.4766
E009	0.2072	0.2145	0.2064	**0.2160**	0.1954	0.1958	0.0943
E010	0.0916	0.0771	0.0753	0.1008	0.1108	**0.1109**	0.1020
E011	0.0698	0.0805	0.0813	**0.1591**	0.0819	0.0845	0.0609
E012	**0.3560**	0.3309	0.3277	0.3150	0.3293	0.3341	0.2858
E013	0.6030	0.6033	0.6096	**0.6188**	0.5872	0.5910	0.5385
E014	0.2008	0.2585	0.2579	**0.2744**	0.2706	0.2694	0.2138
E015	0.1599	0.1583	0.1622	0.1562	0.1795	**0.1795**	0.0964
All	0.2460	0.2484	0.2479	**0.2607**	0.2515	0.2522	0.2095

7 Conclusion

We proposed using the segment-based approach for multimedia event detection in this work. We evaluated our approach by using the state-of-the-art dense trajectories motion feature on the TRECVID MED 2010 and TRECVID MED 2011 datasets. Our proposed segment-based approach outperforms the video-based approach in most cases when using a simple non-overlapping sampling strategy. More interestingly, the results are significantly improved when we using the segment-based approach with an overlapping sampling strategy. Therefore, the effectiveness of our methods on realistic datasets like MEDs is confirmed.

A segment-based approach with an overlapping sampling strategy shows promising results. This suggests the importance of segment localization on the MED performance. Suppose the segment length is fixed, we are interested in determining which segment is the best representative for an event. In this study, we also observed that the detection performance is quite sensitive to the segment-length and it depends on the dataset. The results obtained from the late fusion strategy is quite stable and close the peak performance. This suggests a methodical way to generalize the segment-based approach to other datasets. However, this method is not scalable because it requires a lot of computation costs. Therefore, learning an optimal segment length for each event can be beneficial for an event detection system. This is also an interesting direction for our future study.

Acknowledgments This research is partially funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2013-26-01.

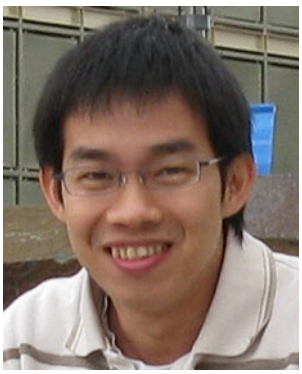
References

1. Chen, M., & Hauptmann, A. (2009). Mosift: recognizing human actions in surveillance videos. In *Computer science department, CMU-CS-09-161*.
2. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (pp. 1–22).
3. Dalal, N., Triggs, B., Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*. Springer.
4. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *International conference on computer vision & pattern recognition* (vol 2, pp. 886–893). INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334. http://lear.inrialpes.fr/pubs/2005/DT05.
5. Duan, L., Xu, D., Chang, S.F. (2012). Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1338–1345). IEEE.
6. Duan, L., Xu, D., Tsang, I.W.H., Luo, J. (2012). Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(9), 1667–1680.
7. Guimarães, S.J.F., Couprie, M., Araújo, A.d.A., Leite, N.J. (2003). Video segmentation based on 2d image analysis. *Pattern Recognition Letters*, *24*(7), 947–957.
8. Hill, M., Hua, G., Natsev, A., Smith, J.R., Xie, L., Huang, B., Merler, M., Ouyang, H., Zhou, M. (2010). Ibm research trecvid-2010 video copy detection and multimedia event detection system. In *NIST TRECVID workshop*. Gaithersburg.
9. Jiang, Y.G., Ngo, C.W., Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on image and video retrieval*, (pp. 494–501).
10. Jiang, Y.G., Zeng, X., Ye, G., Bhattacharya, S., Ellis, D., Shah, M., Chang, S.F. (2010). Columbia-ucf trecvid2010 multimedia event detection: combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID workshop*. Gaithersburg.
11. Jiang, Y., Yuan, J., Yu, G. (2012). Randomized spatial partition for scene recognition. *ECCV*, *2*, 730–743.
12. Kläser, A., Marszałek, M., Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *British machine vision conference* (pp. 995–1004). http://lear.inrialpes.fr/pubs/2008/KMS08.
13. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on computer vision & pattern recognition*. http://lear.inrialpes.fr/pubs/2008/LMSR08.
14. Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2–3), 107–123.
15. Lee, C.H., Soong, F., Juang, B.H. (1988). A segment model based approach to speech recognition. In *International conference on acoustics, speech, and signal processing, 1988. ICASSP-88* (Vol. 1, pp. 501–541). doi:10.1109/ICASSP.1988.196629.
16. Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal Computer Vision*, *60*(2), 91–110.
17. Matsuo, T., & Nakajima, S. (2010). Nikon multimedia event detection system. In *NIST TRECVID workshop*. Gaithersburg.
18. Natarajan, P., Manohar, V., Wu, S., Tsakalidis, S., Vitaladevuni, S.N., Zhuang, X., Prasad, R., Ye, G., Liu, D. (2011). Bbn viser trecvid 2011 multimedia event detection system. In *NIST TRECVID workshop*. Gaithersburg.
19. Natarajan, P., Natarajan, P., Wu, S., Zhuang, X., Vazquez-Reina, A., Vitaladevuni, S.N., Tsourides, K., Andersen, C., Prasad, R., Ye, G., Liu, D., Chang, S., Saleemi, I., Shah, M., Ng, Y., White, B., Gupta, A., Haritaoglu, I. (2012). Bbn viser trecvid 2012 multimedia event detection and multimedia event recounting systems. In *NIST TRECVID workshop*. Gaithersburg, États-Unis.
20. Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal Computer Vision*, *42*(3), 145–175.
21. Oneata, D., Douze, M., Revaud, J., Jochen, S., Potapov, D., Wang, H., Harchaoui, Z., Verbeek, J., Schmid, C., Aly, R., Mcguiness, K., Chen, S., O'Connor, N., Chatfield, K., Parkhi, O., Arandjelovic, R., Zisserman, A., Basura, F., Tuytelaars, T. (2012). AXES at TRECVid 2012: KIS, INS, and MED. In *TRECVID workshop*. Gaithersburg, États-Unis. http://hal.inria.fr/hal-00746874.
22. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M. (2010). Evaluating color descriptors for object and scene recognition. In

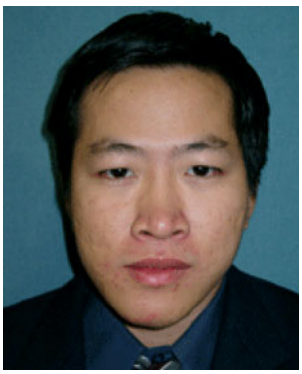
IEEE transactions on pattern analysis and machine intelligence (Vol. 32, pp. 1582–1596).
23. Wang, H., Kläser, A., Schmid, C., Liu, C.L. (2011). Action recognition by Dense Trajectories. In *IEEE conference on computer vision & pattern recognition* (pp. 3169–3176). Colorado Springs. http://hal.inria.fr/inria-00583818/en.
24. Xu, D., & Chang, S.F. (2008). Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1985–1997.
25. Yuan, J., Liu, Z., Wu, Y. (2011). Discriminative video pattern search for efficient action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(9), 1728–1743. doi:10.1109/TPAMI.2011.38.

Sang Phan is a PhD student at The Graduate University for Advanced Studies (SOKENDAI), Japan and a research assistant at the National Institute of Informatics (NII), Japan. He received his BS and MS degrees in 2009 and 2012 from University of Science, Ho Chi Minh City, Vietnam. His research interests include computer vision and multimedia analysis.

Thanh Duc Ngo obtained his BS degrees in Computer Science from the University of Science, Ho Chi Minh City, Vietnam in 2006. He is currently a PhD student at the Department of Informatics, The Graduate University for Advanced Studies (SOKENDAI), Japan. His research interests include pattern recognition, computer vision and multimedia analysis.



Vu Lam is a lecturer and vice dean at the Faculty of Information Technology, University of Science (FIT-HCMUS), Ho Chi Minh City, Vietnam, where he also obtained his BS and MS degrees in 2000 and 2004, respectively. He is currently a PhD student at FIT-HCMUS. He was a visiting researcher at the National Institute of Informatics (NII), Japan, in 2011,2012 and 2013. His research interests include multimedia event detection and violent scene detection.



Son Tran received his Bachelor's degree in science from the Department of Information Technology, the University of Natural Sciences, HCM city, Vietnam in 1997, and his Ph.D's degree in Engineering from the Department of Electronic and Information Science, Toyota Technological Institute, Japan in 2005. Now, he is a lecturer of Faculty of Information Technology, University of Science, Hochiminh city, Vietnam. His research interests include DSP, computer vision, and machine learning.

Duy-Dinh Le received his BS and MS degrees in 1995 and 2001, respectively, from the University of Science, Ho Chi Minh City, Vietnam, and his PhD degree in 2006 from The Graduate University for Advanced Studies (SOKENDAI), Japan. He is currently an assistant professor at the National Institute of Informatics (NII), Japan. His research interests include semantic concept detection, video analysis and indexing, pattern recognition, machine learning, and data mining.

Duc Anh Duong is a professor at the University of Information Technology, Ho Chi Minh City, Vietnam. He obtained his B.Sc and M.Sc in computer science from the University of Ho Chi Minh City in 1990 and 1995, respectively. He received his Ph.D. degree in mathematics from the University of Science, VNU-HCM, Vietnam in 2002. He was a visiting researcher at Japan Advanced Institute of Science and Technology (JAIST), Japan from 2008 to 2009. His research interests include image processing, computer vision and pattern recognition, cryptography and security, geographic information systems, and computer graphics. He is currently the President of University of Information Technology, Chair of Program of Information Technology and Electronics of Ho Chi Minh City, and Chair of Program of Information Security of Ho Chi Minh City, Vietnam. He is a member of the ACM and IEEE.

Shin'ichi Satoh is a professor at the National Institute of Informatics (NII), Japan. He received his BE degree in 1987, ME and PhD degrees in 1989 and 1992 from the University of Tokyo. His research interests include video analysis and multimedia databases. He was a visiting scientist at the Robotics Institute, Carnegie Mellon University, from 1995 to 1997.